Utilizing response times in computerized classification testing

Sie, H.; Finkelman, M.D.; Riley, B.; Smits, N.

DOI
10.1177/0146621615569504
Publication date
2015
Document Version
Final published version
Published in
Applied Psychological Measurement

Article

Utilizing Response Times in Computerized Classification Testing

Applied Psychological Measurement
2015, Vol. 39(5) 389–405

DOI: 10.1177/0146621615569504
apm.sagepub.com
SAGE

Haskell Sie[1], Matthew D. Finkelman[2], Barth Riley[3], and Niels Smits[4]

Abstract

A well-known approach in computerized mastery testing is to combine the Sequential Probability Ratio Test (SPRT) stopping rule with item selection to maximize Fisher information at the mastery threshold. This article proposes a new approach in which a time limit is defined for the test and examinees' response times are considered in both item selection and test termination. Item selection is performed by maximizing Fisher information per time unit, rather than Fisher information itself. The test is terminated once the SPRT makes a classification decision, the time limit is exceeded, or there is no remaining item that has a high enough probability of being answered before the time limit. In a simulation study, the new procedure showed a substantial reduction in average testing time while slightly improving classification accuracy compared with the original method. In addition, the new procedure reduced the percentage of examinees who exceeded the time limit.

Keywords

item response theory, Computerized Classification Testing, response time, lognormal model, Sequential Probability Ratio Test

Introduction

In Computerized Classification Testing (CCT), examinees are classified into one of multiple proficiency groups based on their responses to test items. These item responses serve as an indicator of where the examinees' abilities are on the latent trait continuum relative to the mastery thresholds of the test. CCT allows the number of test items to differ for each examinee: unless there is a requirement that the same number of items be given to all examinees, the test for any particular examinee can be terminated as soon as there is enough evidence to make a

[1]American Institutes for Research, Washington, DC, USA
[2]Tufts University School of Dental Medicine, Boston, MA, USA
[3]University of Illinois at Chicago, USA
[4]Vrije Universiteit Amsterdam, The Netherlands

Corresponding Author:
Haskell Sie, American Institutes for Research, 1000 Thomas Jefferson Street NW, Washington, DC 20007-3835, USA.
Email: hsie@air.org

classification decision. In the particular context of CCT with only one mastery threshold θ_c and two proficiency groups (which will be the focus of this article), the test can be terminated and the examinee passes (or fails) the test if evidence collected based on test responses indicates that his or her latent ability is above (or below) the threshold for mastery.

In any implementation of CCT, the item selection method and stopping rule play a vital role. Selection of test items is commonly done based on the maximum Fisher information principle. Two variations of this framework exist in CCT: choosing items maximizing information at an interim estimate of the examinee's ability and choosing items maximizing information at the mastery threshold. Regarding the choice of a stopping criterion, the Sequential Probability Ratio Test (SPRT) has received much attention in the literature (Spray, 1993; Thompson, 2011; Wald, 1947; Weissman, 2007). This stopping rule stems from Wald's (1947) treatise on sequential analysis, in which the author proposed that inferences be based on a likelihood ratio (LR) statistic that is updated after each observation is obtained. With the SPRT used as the stopping criterion in CCT, Spray and Reckase (1994) compared the performance of an item selection method that maximizes information at the mastery threshold and one that maximizes information at the examinee's true ability. It was found that the former outperformed the latter in terms of Average Test Length (ATL). While item selection maximizing information at the examinee's most recent ability estimate was not used in their study, Spray and Reckase argued that the ATL of said method is expected to be even longer than that obtained when the examinee's true ability is used.

More recent advances in computerized testing have utilized data on the amount of time that each examinee spends on each test item. Such response time (RT) data provide an additional useful piece of information regarding the test-takers as well as the test items (van der Linden, 2008). Ample applications of RT modeling have been discussed in the literature. Their applications in the framework of computerized adaptive testing (CAT) include detection of aberrant responses (van der Linden & van Krimpen-Stoop, 2003), detection of advanced item knowledge (Meijer & Sotaridona, 2006), control of speededness (van der Linden, 2009; van der Linden, Scrams, & Schnipke, 1999; van der Linden & Xiong, 2013), and improvement of item selection rules (Fan, Wang, Chang, & Douglas, 2012; van der Linden, 2008). In particular, the study by Fan et al. (2012) showed that by taking into account examinees' expected RTs when choosing test items, the average time needed by examinees to complete the test can be substantially reduced, albeit with a small loss in estimation accuracy. The study, however, did not assess performance of the proposed item selection method in the context of CCT.

The current research aims at extending the work of Fan et al. (2012) by developing a CCT procedure that is suited to tests with time restrictions. The procedure includes both an item selection method and a stopping rule that take the element of time into consideration. The goal of taking RTs into account when choosing test items is to produce a shorter testing time while maintaining high classification accuracy. Reducing the testing time (as opposed to focusing on reducing the number of items administered) is desirable in many applications of CCT. For example, in low-stakes educational tests used for diagnostic purposes, or in health and psychological assessments, tests that consist of more items with shorter total duration may be preferable to tests that have fewer items but require more total time. Longer test duration might lead to test-takers becoming less focused or less motivated to answer the items.

In addition to producing a shorter testing time, it is important to take into account the fact that many tests have a specified time limit, the surpassing of which is undesirable. When such a time limit has been reached, the assessment can either be halted immediately (in which case the final item is not completed), or the examinee may be allowed to finish the final item before examination is ceased. In some testing applications (e.g., low-stakes diagnostic assessments), the latter rule may be preferable so that the examinee is allowed to complete all items that have

been presented to him or her. In this case, it is acceptable for some examinees to exceed the given time limit, as long as a specified percentage of examinees do not exceed it. Motivated by the need of a stopping rule that takes into account the time limit for the CCT, a modification to the SPRT that will control the percentage of examinees who fall within the desired time limit is proposed. In what follows, the modified stopping rule is referred to as the ''time-limited SPRT stopping rule.''

The remainder of the article is organized as follows. The next section briefly reviews commonly used item response theory (IRT) and RT models, as well as the SPRT. Following that, the two item selection methods as well as the time-limited SPRT stopping rule used in this article are explained. Simulation results are then presented, followed by concluding remarks and directions for future work.

Theoretical Background

Models for Test Responses and RT

To conduct CCT, it is necessary to formalize the relationship between the latent trait being measured (which is often referred to as ''ability'' in educational assessments) and the item responses. When items are dichotomous, such a relationship is commonly characterized via the three-parameter logistic (3PL) model (Birnbaum, 1968). Defining U_i as a Bernoulli random variable that takes the value 1 if item i is answered correctly and 0 otherwise, the probability of an examinee with ability θ answering item i correctly is given by,

$$P_i(\theta) \equiv P(U_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp\{-a_i(\theta - b_i)\}}, \tag{1}$$

where a_i is the discrimination parameter, b_i is the difficulty parameter, and c_i is the guessing parameter of item i.

Using the model in Equation 1, the likelihood function upon observing an examinee's response to item i is given by

$$L(u_i|\theta) = \{P_i(\theta)\}^{u_i}\{Q_i(\theta)\}^{1-u_i}, \tag{2}$$

where $Q_i(\theta) = 1 - P_i(\theta)$ and a lowercase u_i is used to denote a realization of the random variable U_i. The likelihood function upon observing the responses to n test items is,

$$L_n \equiv L(u_1, \ldots, u_n|\theta) = \prod_{i=1}^{n} L(u_i|\theta) = \prod_{i=1}^{n} \{P_i(\theta)\}^{u_i}\{Q_i(\theta)\}^{1-u_i}, \tag{3}$$

which is a consequence of the assumption that all test responses are independent given θ (i.e., the local independence assumption).

The likelihood function in Equation 2 can be used to define the Fisher information for item i as,

$$\mathrm{FI}_i(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\{\log L(U_i|\theta)\}\right] = \frac{Q_i(\theta)\{P_i(\theta) - c_i\}^2}{P_i(\theta)(1 - c_i)^2}a_i^2, \tag{4}$$

and similarly, the Fisher information of a test of n items can be derived from Equation 3 as,

$$\mathrm{FI}^{(n)}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\{\log L_n\}\right] = \sum_{i=1}^{n}\mathrm{FI}_i(\theta). \tag{5}$$

As mentioned earlier, the use of computers in administering tests provides not only data from examinees' test responses but also their RT. In this article, the lognormal model of van der Linden (2006) is adopted in modeling RT data. For an examinee who works at a constant speed τ, the time T_i spent on answering item i is a random variable with probability density function given by,

$$f(t_i) \equiv f_{T_i}(t_i|\tau) = \frac{\alpha_i}{t_i\sqrt{2\pi}}\exp\left[-\frac{\alpha_i^2}{2}\{\ln t_i - (\beta_i - \tau)\}^2\right], \tag{6}$$

where α_i is the discrimination parameter of item i with respect to speed, β_i is the time-intensity parameter of item i, and a lowercase t_i is used to denote a realization of the random variable T_i. As α_i is inversely proportional to the standard deviation of the RT distribution, a larger value means that the person's RT to item i would be less dispersed, if replications were possible. The parameter β_i affects the mean of the RT distribution: The larger the β_i, the larger the amount of time an examinee will spend on item i, on average. Bayesian estimation of all model parameters in Equation 6 can be performed with Markov Chain Monte Carlo (MCMC) as in van der Linden (2007).

If it is of interest to model the correlation between examinee ability (θ) and speed (τ) in the population, the more complete modeling framework as in van der Linden (2007) can be used. This hierarchical framework uses the models in Equations 1 and 6 for test responses and RTs, respectively. In addition, it specifies that the vector $\xi = \begin{pmatrix}\theta\\ \tau\end{pmatrix}$ of person parameters has the bivariate normal joint distribution, denoted N_2; that is,

$$\boldsymbol{\xi} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{7}$$

where the mean vector,

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_\theta\\ \mu_\tau\end{pmatrix}, \tag{8}$$

and the covariance matrix,

$$\boldsymbol{\Sigma} = \begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta\tau}\\ \sigma_{\theta\tau} & \sigma_\tau^2\end{pmatrix}. \tag{9}$$

To obtain identifiability, further constraints need to be imposed on the model, for example, by setting $\mu_\theta = 0$, $\sigma_\theta^2 = 1$, and $\mu_\tau = 0$. Hereafter, it will be assumed that $\sigma_{\theta\tau}$, σ_τ^2, and all item parameters have been estimated with enough precision to be treated as known constants.

The SPRT

To achieve the goal of making efficient classification decisions in CCT, a statistical hypothesis testing problem of the form $H_0 : \theta < \theta_c$ versus $H_1 : \theta \geq \theta_c$ is first formulated. If the null hypothesis is rejected, evidence is available based on the examinee's test responses that his or her ability is at or beyond the mastery threshold of the test and thus that he or she should pass the test. In the psychometric literature, the above testing problem is usually modified to the problem of testing $H_0 : \theta \leq \theta_-$ versus $H_1 : \theta \geq \theta_+$, where (θ_-, θ_+) is a small neighborhood around θ_c,

called the *indifference region*. A typical width for the indifference region is usually about 0.4 points on the ability scale; that is, a width of 0.2 in each direction (Eggen, 1999). For examinees with abilities inside the indifference region, it is assumed that the test administrator is indifferent as to which proficiency group they are eventually assigned to; their abilities are sufficiently close to the mastery threshold that neither classification decision is considered a mistake. Furthermore, to test the above modified composite hypotheses subject to error constraints,

$$P_\theta(\text{reject } H_0) \leq \alpha \quad \text{for } \theta \leq \theta_-, \tag{10}$$

and

$$P_\theta(\text{not reject } H_0) \leq \beta \text{ for } \theta \geq \theta_+, \tag{11}$$

surrogate hypotheses of the form,

$$H_0 : \theta = \theta_- \text{ versus } H_1 : \theta = \theta_+, \tag{12}$$

are usually used. Here, $\alpha = P_{\theta_-}(\text{reject } H_0)$ denotes the Type I error probability (i.e., the probability of incorrectly stating that the examinee has mastered the test when he or she has not) and $\beta = P_{\theta_+}(\text{not reject } H_0)$ denotes the Type II error probabilities (i.e., the probability of incorrectly stating that the examinee has not mastered the test when he or she has). Note that in the notation, α and β (the Type I and Type II error rates) are distinguished from α_i and β_i in Equation 6 (the discrimination parameter with respect to speed and the time-intensity parameter) by the subscripts of the latter.

For the simple hypotheses in Equation 12, the SPRT is used to determine when a classification decision in favor of either H_0 or H_1 can be made, at which point, the CCT can be terminated. The stopping time can be mathematically formulated as,

$$\varphi = \inf\{m \geq 1 | \lambda_m \leq B \text{ or } \lambda_m \geq A\}, \tag{13}$$

where inf() denotes the infimum of a set, A and B are stopping boundaries with $0 < B < 1 < A < \infty$, and

$$\lambda_m = \frac{L(u_1, \ldots, u_m|\theta_+)}{L(u_1, \ldots, u_m|\theta_-)}, \tag{14}$$

is a LR statistic computed after the mth item is answered. When stopping occurs, the null hypothesis in Equation 12 will be rejected if $\lambda_\varphi \geq A$, and not rejected if $\lambda_\varphi \leq B$. The choice of the stopping boundaries A and B is determined by the desired Type I and Type II error probabilities defined earlier. Wald (1947) proposed using $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$ to approximately achieve the desired α and β values. In the usual application of SPRT in CCT, a maximum number of test items, say $n_{\max}$, is used as a constraint. When a forced truncation occurs after $n_{\max}$ test items are administered, the SPRT's classification decision is that the examinee passes the test if and only if $\lambda_{n_{\max}} \geq C$, where C, the crossing boundary at truncation, is commonly set at 1 if H_0 and H_1 are to be treated symmetrically (Spray & Reckase, 1996).

Using RT to Modify Item Selection Criteria in CCT

When test items are selected based on the maximum Fisher information principle, the goal is to maximize the Fisher item information in Equation 4 in the current context of tests measuring only a unidimensional ability. Suppose that $k - 1$ test items have been administered from a pool

of N items, and that S_{k-1} and $R_{k-1} = \{1, \ldots, N\} \cap S_{k-1}^{C}$ denote the collection of test items that have and have not been administered, respectively (here, A^C means the complement of the set A). Then, maximizing Fisher item information at an interim ability estimate $\hat{\theta}_{k-1}$ implies that the kth test item is chosen according to

$$i_k = \underset{j \in R_{k-1}}{\arg\max}\, \mathrm{FI}_j\left(\hat{\theta}_{k-1}\right). \tag{15}$$

However, maximizing Fisher item information at the mastery threshold implies that the kth test item will be chosen according to

$$i_k = \underset{j \in R_{k-1}}{\arg\max}\, \mathrm{FI}_j(\theta_c). \tag{16}$$

Recently, Fan et al. (2012) introduced a CAT item selection method that chose items having the ''maximum information per time unit.'' Specifically, they proposed choosing the kth test item to be

$$i_k = \underset{j \in R_{k-1}}{\arg\max} \frac{\mathrm{FI}_j\left(\hat{\theta}_{k-1}\right)}{E\left(T_j|\hat{\tau}_{k-1}\right)}, \tag{17}$$

where $\hat{\theta}_{k-1}$ and $\hat{\tau}_{k-1}$ are any interim estimates of the examinee's ability and speed parameters, respectively, and $E(T_j|\hat{\tau}_{k-1}) = \exp\left(\beta_j - \hat{\tau}_{k-1} + \frac{1}{2\alpha_j^2}\right)$ following the lognormal model in Equation 6 for the distribution of T_j (Fan et al., 2012). Using maximum likelihood to estimate both θ and τ, Fan et al. showed that estimation accuracy only differed slightly between the original maximum information criterion in Equation 15 and that in Equation 17, but that using the latter substantially reduced the average time needed by examinees to complete the test.

For the current study, the ''maximum information per time unit'' framework introduced by Fan et al. (2012) is extended to the current context of CCT. It is seen in Equation 17 that the numerator of the item selection criterion is evaluated at an interim ability estimate $\hat{\theta}_{k-1}$, which, for CCT, results in higher ATL than Fisher information at the mastery threshold (Spray & Reckase, 1994). Therefore, the new item selection method combines the use of Fisher item information at the mastery threshold and the use of expected RT, leading to the following item selection method: Choose the kth test item to be

$$i_k = \underset{j \in R_{k-1}}{\arg\max} \frac{FI_j(\theta_c)}{E\left(T_j|\hat{\tau}_{k-1}\right)}. \tag{18}$$

Given that the item selection method in Equation 18 takes RT into account, it should be coupled with a stopping rule that also considers RT. To this end, the time-limited SPRT stopping rule is introduced.

Using RT to Modify the SPRT Stopping Rule in Low-Stakes CCT

In ''The SPRT'' section, the SPRT was introduced as a stopping criterion in CCT. It was mentioned that the SPRT is typically truncated when a certain maximum number of test items $n_{\max}$ has been reached. In this section, another truncation procedure will be introduced for situations in which there is a desired time limit $t_{\max}$ that is imposed on the test. To the best of the authors' knowledge, the issue of a timed CCT has never been discussed in the literature. Nearly all of the

literature on CCT has discussed termination procedures with regard to early stopping and/or the use of a maximum test length that is set as a constraint. Although setting a maximum test length seems to be the approach taken by most studies on CCT in the literature, it might not always be the most practical approach. If testing times are not considered, an unduly high percentage of examinees could exceed the desired time limit. This in turn would require more resources for that particular testing window, such as availability of computer stations. In addition, for many applications in clinical settings, it is desired that the respondent burden be minimized (Lohr, 2002). With such considerations in mind, it becomes clear that sound procedures are currently still needed for effective administration of a timed CCT.

In a timed CCT, an automated termination procedure must be available that can be readily invoked when the specified time limit is reached. The following modification has been proposed to the regular SPRT: After the kth test item is administered (using one of the methods given in Equation 15, 16, 17, or 18) and answered, define λ_k, A, and B as in ''The SPRT'' section, and $t_k^* = t_{\max} - \sum_{i=1}^{k} t_i$. Stop testing if

$$\lambda_k \geq A \text{ or } \lambda_k \leq B, \tag{19}$$

if

$$B < \lambda_k < A \text{ and } t_k^* \leq 0, \tag{20}$$

or if

$$B < \lambda_k < A,\ t_k^* > 0 \text{ and } E_k = \varnothing, \tag{21}$$

where $E_k = \left\{ j \in R_k | P\left(T_j > t_k^*\right) \leq \gamma \right\}$ is the collection of test items that have not yet been administered and have a probability smaller than γ of requiring more time to be answered than there is time remaining. Equation 19 is the regular ''early stopping'' condition of the SPRT: The test is terminated when a classification decision can already be made with enough certainty based on the specified Type I and Type II error probabilities. Equation 20 truncates the SPRT by terminating the test if the time limit has been met or exceeded following the most recent item, even though a classification decision based solely on the value of λ_k cannot yet be made. Equation 21 attempts to prevent the examinee from exceeding the time limit: It terminates the test if all unused items in the pool have probability greater than a specified constant (namely, γ) of requiring more time than remains in the test. A similar approach to choose test items with expected RT not exceeding the remaining time limit was used in van der Linden (2009) for CAT that reports an estimate of examinee's ability instead of a classification decision. It should be noted that the stopping criteria in Equations 19 to 21 do not guarantee that no examinee will exceed the time limit. There could be instances whereby the set E_k is not empty, but that the selected item from E_k takes a longer time to be answered than the remaining time, a consequence of the examinee's RT being a random variable. However, because the percentage of examinees who exceed the time limit depends on the value of γ, this percentage can be controlled through a judicious choice of the constant. A note has been made that when Equations 19 to 21 are used as a stopping rule, it is desired that the test item to be administered next will not require more time than remains. Therefore, in the item selection, the set of possible items to come from the set E_k instead of R_k has been restricted.

In Equation 21, it is to be evaluated whether each unused item $j \in R_k$ is eligible to be included in the set E_k by means of whether $P(T_j > t_k^*) \leq \gamma$. This probability can be computed

conditioning on the previous RTs up to and including the kth item. Letting $t_k = (t_1, \ldots, t_k)$, we have,

$$P\left(T_j > t_k^* | \boldsymbol{t}_k\right) = \int P\left(T_j > t_k^* | \tau\right) f(\tau | \boldsymbol{t}_k) d\tau. \tag{22}$$

In Equation 22, the first term in the integrand can be evaluated using the cumulative distribution function of T_j based on the lognormal model in Equation 6, whereas the second term is the Bayesian posterior distribution of τ given by,

$$f(\tau | \boldsymbol{t}_k) \sim N\left(\frac{\sigma_\tau^{-2}\mu_\tau + \sum_{i=1}^{k} \alpha_i^2(\beta_i - \ln t_i)}{\sigma_\tau^{-2} + \sum_{i=1}^{k} \alpha_i^2}, \frac{1}{\sigma_\tau^{-2} + \sum_{i=1}^{k} \alpha_i^2} \right), \tag{23}$$

for a prior mean μ_τ and a prior variance σ_τ^2 of τ (van der Linden, 2008). Here and in what follows, the notation $N(\mu, \sigma^2)$ is used to denote the normal distribution with mean μ and variance σ^2. The integral in Equation 22 can then be evaluated using any integration package in standard statistical software. In particular, in the simulation, the integrate() function in R was used, setting -4 and $+4$ as the upper and lower integration bounds, respectively, which reflects the common range of τ values.

Next, the question of how to make a classification decision using the time-limited SPRT stopping rule is addressed. If the test is terminated due to a classification decision being made early (see Equation 19), the usual SPRT decision is employed at the occurrence of early stopping; that is, the null hypothesis in Equation 12 will be rejected if $\lambda_\varphi \geq A$, and not rejected if $\lambda_\varphi \leq B$. However, if the test is terminated due to a time consideration (see Equation 20 or 21), an examinee passes the test if and only if $\lambda_{n^*} \geq C$, where n^* is the last item answered by the examinee before the test is terminated and C is the crossing boundary at truncation.

A step-by-step guide regarding the implementation of the procedure is given in the online appendix.

Simulation

The previous section described motivation for utilizing additional information obtained from examinees' RTs to modify both the maximum information item selection method and the SPRT stopping rule in CCT. In this section, results of a simulation study are presented that illustrate relative performance of the item selection methods in Equations 16 and 18 under the time-limited SPRT stopping rule. In addition, these methods are compared with the standard practice in CCT that combines the item selection method in Equation 16 and the regular SPRT.

Design

Following Fan et al. (2012), the 3PL model was used in the simulation with an item pool consisting of 500 items with IRT a parameters from the $U(1.0, 2.5)$ distribution, where $U(a, b)$ denotes the uniform distribution in (a, b), b parameters from $N(0, 1)$, and c parameters from $\beta(2, 10)$. The discrimination parameters with respect to speed (i.e., α) were generated from the $U(1, 3)$ distribution whereas the time-intensity parameters were generated in two different ways: (i) assuming no correlation with IRT b parameters, the β parameters were generated from the $U(3, 5)$ distribution (van der Linden, 2008) and (ii) assuming a 0.65 correlation with IRT b

parameters (van der Linden et al., 1999), the β parameters were generated by sampling a separate value from its conditional distribution given b; that is, $N(\mu_\beta + \rho_{b\beta}\sigma_\beta/\sigma_b(b-\mu_b), \sigma_\beta^2 - \rho_{b\beta}^2\sigma_\beta^2,,$ under the assumption of a bivariate normal joint distribution for $\begin{pmatrix} b \\ \beta \end{pmatrix}$, similar to Equations 7 to 9 with θ replaced by b and τ replaced by β. For approach (ii), $\mu_\beta = 4$ and $\sigma_\beta^2 = 1/3$ were used to give the same mean and variance of the β parameters as in approach (i).

Five hundred examinees were simulated at each of 25 evenly spaced true ability values from −3.0 to +3.0 with an increment of 0.25. Similar to the time-intensity parameters, examinees' speed parameters were also generated in two different ways: (i) assuming no correlation with θ, the τ parameters were generated from the $N(0, 0.24^2)$ distribution and (ii) assuming a 0.59 correlation with θ (van der Linden, 1999), the τ parameters were generated by sampling a separate value from its conditional distribution given θ; that is, $N(\mu_\tau + \rho_{\theta\tau}\sigma_\tau/\sigma_\theta(\theta-\mu_\theta), \sigma_\tau^2 - \rho_{\theta\tau}^2\sigma_\tau^2,$ under the assumption of a bivariate normal joint distribution for $\begin{pmatrix} \theta \\ \tau \end{pmatrix}$, as in Equations 7 to 9. For approach (ii), $\mu_\tau = 0$ and $\sigma_\tau^2 = 0.24^2$ were used to give the same mean and variance of the τ parameters as in approach (i). The two approaches for generating β parameters as well as τ parameters led to four correlation structures.

Similar to Spray and Reckase (1994), three locations of the mastery threshold were used ($\theta_c = -1$, $\theta_c = 0$, and $\theta_c = 1$). Regardless of location of the mastery threshold, an indifference region of width = 0.4 was constructed (i.e., a width of 0.2 in each direction), similar to one simulation in Eggen (1999). The three locations of the mastery threshold and four correlation structures were crossed, leading to 12 study conditions. For the time-limited SPRT stopping rule, $t_{max} = 900$ s was set as the time limit. In addition, both the nominal Type I and Type II error probabilities were set at 0.05, leading to crossing boundaries $A = 19$ and $B = 1/19$. The crossing boundary at truncation was set at $C = 1$, and the constant γ was set at 0.05. For the regular SPRT stopping rule, the maximum number of test items to be administered before truncation was chosen such that the resulting proportion of correct decision (PCD) was as close as possible to those from other simulations herein that used the time-limited SPRT stopping rule.

Test items were selected to either maximize information or information per time unit at the mastery threshold following Equation 16 or 18, respectively. To avoid having to compute the Fisher information multiple times for each item, an information matrix was constructed prior to the simulation. This information matrix contained the Fisher information of each item at each of the three mastery thresholds. To standardize results across methods, the same response and RT were used whenever more than one method administered the same item to the same examinee. This was done by constructing two matrices prior to the simulation, one containing each examinee's response to each item and another containing each examinee's RT to each item, both randomly generated according to the models in Equations 1 and 6, respectively.

Results

In what follows, the item selection method that maximizes information at the mastery threshold (see Equation 16) will be referred as M1 and another that maximizes information per time unit at the mastery threshold (see Equation 18) as M2. The two item selection methods are first compared based on their ATL. Figure 1 displays the ATL in all study conditions by means of a 4 × 3 plot. Different columns represent different locations of the mastery threshold ($\theta_c = -1$ in the first column, $\theta_c = 0$ in the second column, and $\theta_c = 1$ in the third column), whereas different

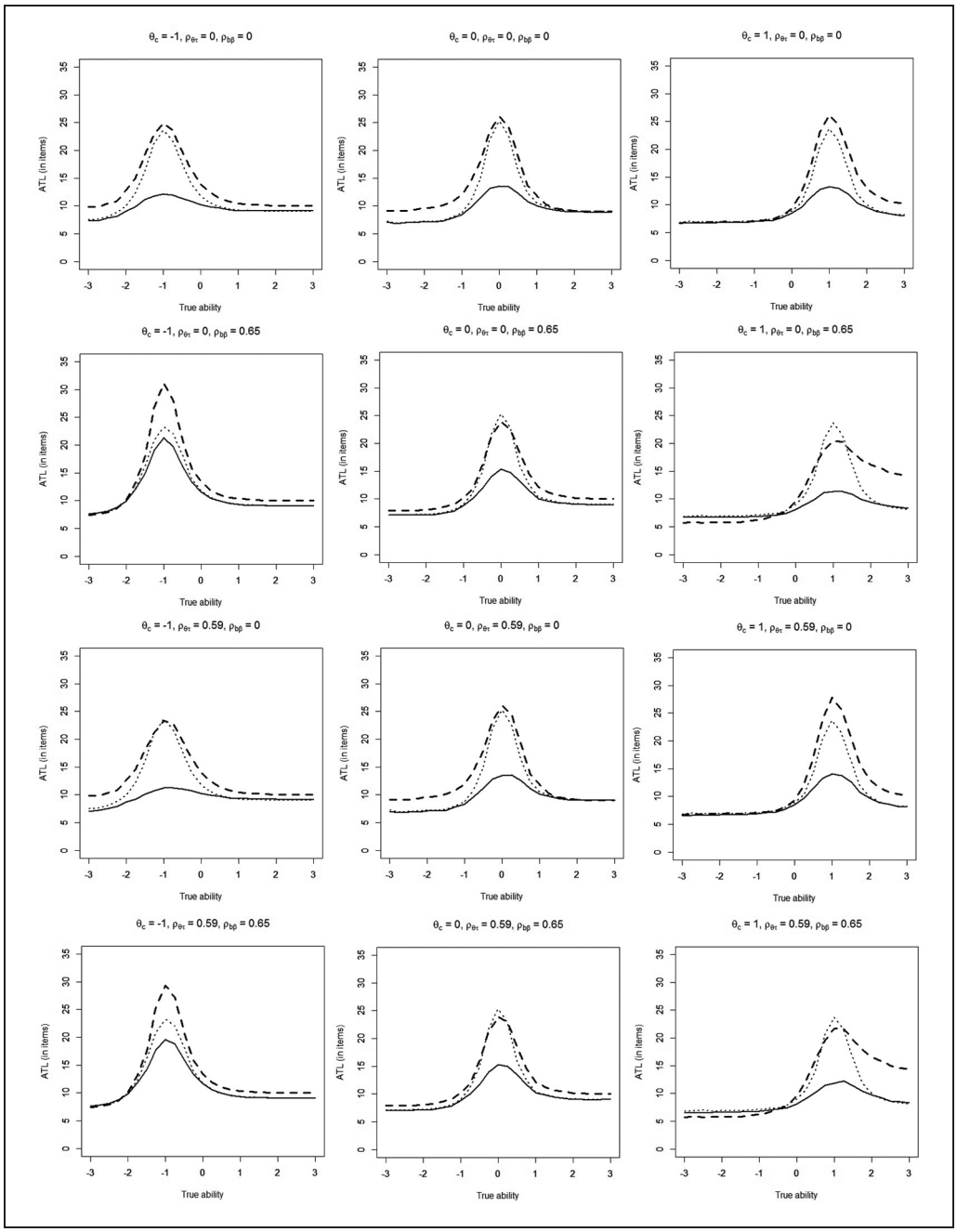

Figure 1. ATL comparison between M1 under the regular SPRT (dotted line), M1 under the time-limited SPRT (solid line), and M2 under the time-limited SPRT (dashed line).

Note. Different columns represent different locations of the mastery threshold ($\theta_c = -1, 0, 1$ in the first, second, and third columns, respectively). Different rows represent different correlation structures (from the first to the fourth row: $\rho_{\theta\tau} = 0$ and $\rho_{b\beta} = 0$, $\rho_{\theta\tau} = 0$ and $\rho_{b\beta} = 0.65$, $\rho_{\theta\tau} = 0.59$ and $\rho_{b\beta} = 0$, $\rho_{\theta\tau} = 0.59$ and $\rho_{b\beta} = 0.65$). ATL = Average Test Length; SPRT = Sequential Probability Ratio Test.

rows represent different correlation structures: The first row displays results when $\rho_{\theta\tau} = 0$ and $\rho_{b\beta} = 0$, the second row when $\rho_{\theta\tau} = 0$ and $\rho_{b\beta} = 0.65$, the third row when $\rho_{\theta\tau} = 0.59$ and $\rho_{b\beta} = 0$, and the fourth row when $\rho_{\theta\tau} = 0.59$ and $\rho_{b\beta} = 0.65$. In all of the plots, the dotted line represents results when M1 was used with the regular SPRT stopping rule, the solid line represents results when M1 was used with the time-limited SPRT stopping rule, and the dashed line represents results when M2 was used with the time-limited SPRT stopping rule. For M1 coupled with the regular SPRT stopping rule, the maximum number of administered items before truncation was set at 25 items when $\theta_c = -1$, at 29 items when $\theta_c = 0$, and 26 items when $\theta_c = 1$. As mentioned earlier, these maximum numbers of test items were chosen in order that the resulting PCDs be as close as possible to those obtained when M1 was coupled with the time-limited SPRT stopping rule.

In general, Figure 1 shows that across all 12 conditions, the test administered under M2 had a higher ATL than tests administered under M1. When the item selection method M1 was coupled with the time-limited SPRT stopping rule, it had a significantly lower ATL than the corresponding test administered under M2 coupled with the same stopping rule, especially around the mastery threshold. The standard practice in CCT that uses M1 to select test items coupled with the regular SPRT stopping rule generally yielded ATL slightly lower than those for tests administered using item selection method M2 coupled with the time-limited SPRT stopping rule, but higher than those for tests with the same item selection method M1 but with the time-limited SPRT stopping rule. This shows that, controlling for the same item selection method M1, the time-limited SPRT stopping rule yielded lower ATL than the regular SPRT stopping rule.

It is also seen in Figure 1 that the correlation between item *b* and β parameters affected ATL more than the correlation between person θ and τ parameters did, especially when the time-limited SPRT stopping rule was used. With this stopping rule, similar ATL was observed for each item selection method and for each ability level when $\rho_{b\beta}$ was held constant but $\rho_{\theta\tau}$ varied. When $\rho_{\theta\tau}$ was held constant but $\rho_{b\beta}$ varied, however, the observed ATL was different at many ability levels within a given item selection method. For example, with $\theta_c = -1$ and $\rho_{\theta\tau} = 0$, ATL for M1 with the time-limited SPRT stopping rule ranged from 7.27 to 12.13 items when $\rho_{b\beta} = 0$, but the range was 7.61 to 21.29 items when $\rho_{b\beta} = 0.65$. When coupled with the time-limited SPRT stopping rule, the item selection method M2 generally had a higher ATL than M1.

Due to the method by which the maximum number of items was chosen, PCDs of tests that selected items based on the method M1 coupled with the regular SPRT stopping rule matched those of tests with the same item selection method coupled with the time-limited SPRT stopping rule. Therefore, only PCDs of tests, where the time-limited SPRT stopping rule was used, will be discussed. Results pertaining to PCD under the time-limited SPRT stopping rule are presented in Table 1 for conditions, where $\rho_{\theta\tau} = \rho_{b\beta} = 0$. PCD results for other conditions are available in Tables 3 to 5 in the online appendix. When $\rho_{\theta\tau} = \rho_{b\beta} = 0$, PCDs of tests administered under M2 were generally higher than those under M1. When the mastery threshold was $\theta_c = -1$, there were some ability levels at which the PCD of M1 was higher than that of M2, but the difference never exceeded 1.2% regardless of the values of $\rho_{\theta\tau}$ and $\rho_{b\beta}$. For most ability levels, M2 had higher PCD than M1, with a PCD advantage up to 5.4%. When the mastery threshold was $\theta_c = 0$ and $\rho_{\theta\tau} = \rho_{b\beta} = 0$, the PCD of M1 was higher than that of M2 at only one ability level ($\theta = -1.5$) by 0.2%. For all other ability values and correlation structures, the PCD of M2 was always higher than that of M1, with a difference of up to 9.6%. In addition, the PCD of M2 was greater than or equal to that of M1 for all ability levels when $\theta_c = 1$. Depending on $\rho_{\theta\tau}$ and $\rho_{b\beta}$, the PCD difference between M1 and M2 ranged from 0% to 4.4% in favor of M2.

From Table 1, the Type I and Type II error probabilities in the simulation can also be observed. Based on the ability values used in the study, it is seen that the error rates under both

Table 1. PCD Under M1 and M2 With the Time-Limited SPRT Stopping Rule, Given at Various Ability Levels When $\rho_{\theta\tau} = \rho_{b\beta} = 0$.

	$\theta_c = -1$		$\theta_c = 0$		$\theta_c = 1$	
Ability	M1	M2	M1	M2	M1	M2
−3.00	1.000	1.000	1.000	1.000	1.000	1.000
−2.75	1.000	1.000	1.000	1.000	1.000	1.000
−2.50	1.000	1.000	1.000	1.000	1.000	1.000
−2.25	1.000	1.000	1.000	1.000	1.000	1.000
−2.00	0.998	1.000	1.000	1.000	1.000	1.000
−1.75	0.990	0.996	1.000	1.000	1.000	1.000
−1.50	0.942	0.974	1.000	0.998	1.000	1.000
−1.25	0.782	0.836	0.998	1.000	1.000	1.000
−1.00	0.510	0.502	0.998	1.000	1.000	1.000
−0.75	0.816	0.834	0.990	1.000	1.000	1.000
−0.50	0.944	0.974	0.964	0.988	1.000	1.000
−0.25	0.994	0.996	0.782	0.878	1.000	1.000
0.00	0.998	1.000	0.516	0.530	1.000	1.000
0.25	1.000	1.000	0.840	0.884	0.994	0.998
0.50	1.000	1.000	0.968	0.984	0.960	0.994
0.75	1.000	1.000	0.990	1.000	0.820	0.840
1.00	1.000	1.000	0.998	1.000	0.502	0.512
1.25	1.000	1.000	1.000	1.000	0.850	0.894
1.50	1.000	1.000	1.000	1.000	0.968	0.990
1.75	1.000	1.000	1.000	1.000	0.994	0.998
2.00	1.000	1.000	1.000	1.000	0.998	1.000
2.25	1.000	1.000	1.000	1.000	1.000	1.000
2.50	1.000	1.000	1.000	1.000	1.000	1.000
2.75	1.000	1.000	1.000	1.000	1.000	1.000
3.00	1.000	1.000	1.000	1.000	1.000	1.000

Note. PCD = proportion of correct decision; SPRT = Sequential Probability Ratio Test.

item selection methods are higher than the nominal values only at ability points inside the indifference region as well as those closest to both ends of the indifference region. At all other ability points under study, the error rates are either very close to or lower than the nominal values. When $\theta_c = -1$, the indifference region ranges from -1.2 to -0.8. For an ability of -1.25, the Type I error rate is higher than the nominal value (0.218 vs. 0.050) under M1. In addition, the Type II error rate is higher than the nominal value for an ability of -0.75 (0.184 vs. 0.050). When $\theta_c = 0$, the indifference region ranges from -0.2 to $+0.2$. It is seen that the Type I error rate is higher than the nominal value for an ability of -0.25 (0.218 vs. 0.050) and that the Type II error rate is higher than the nominal value for an ability of $+0.25$ (0.160 vs. 0.050). When $\theta_c = 1$, the indifference region ranges from $+0.8$ to $+1.2$. For this condition, the Type I error rate is higher than the nominal value for an ability of $+0.75$ (0.180 vs. 0.050) and the Type II error rate is higher than the nominal value for an ability of $+1.25$ (0.150 vs. 0.050). These numbers are based on the item selection method M1 but similar results were observed under M2. Such results are expected due to the nature of the stopping rule. It is well known in the sequential testing literature that error rates will increase if the test is truncated. In previous studies where truncation occurs after a certain number of test items has been administered, the Type I and Type II error rates were also higher than the nominal values (Finkelman, 2008; Thompson, 2011). In the present study, truncation occurred based on a time consideration, but the same phenomenon of increased error rates was observed.

Next, average testing time for the tests is discussed. Figure 2 indicates that in general, using the time-limited SPRT stopping rule significantly reduced testing time from the standard practice in CCT where the regular SPRT stopping rule is used in conjunction with the item selection method M1. With the time-limited SPRT stopping rule, modifying the item selection method from M1 to M2 provides further reduction to testing time, especially at the two tails of the ability distributions. This was a pattern consistent regardless of the location of the mastery threshold or the correlation structure between item parameters or person parameters. In most conditions considered, testing time for examinees with abilities near the mastery threshold was at least twice as long under the standard practice in CCT as when the time-limited SPRT stopping rule was used, whether it was coupled with the item selection method M1 or M2. The results in Figure 2 are displayed with the same arrangement as in Figure 1.

Under the time-limited SPRT stopping rule, it is seen in Figure 2 that reduction in testing time obtained by using the item selection method M2 instead of M1 was largest when $\rho_{b\beta} = 0$. For example, when $\theta_c = -1$ and $\rho_{\theta\tau} = \rho_{b\beta} = 0$, testing time was reduced by an average of 5.24 min across all ability levels. For the same location of mastery threshold and the same value of $\rho_{\theta\tau}$ but with $\rho_{b\beta} = 0.65$, the reduction in testing time averaged at 2.11 min across all ability levels. In addition, it is seen that testing time for examinees with abilities near the mastery threshold was also shorter under M2 than under M1.

To illustrate the ability of the time-limited SPRT stopping rule to control the percentage of examinees who exceed the time limit, Table 2 presents those percentages when item selection method M1 was coupled with the regular SPRT stopping rule, when item selection method M1 was coupled with the time-limited SPRT stopping rule, as well as when item selection method M2 was coupled with the time-limited SPRT stopping rule. For each combination of item selection method and stopping rule, results are displayed for all values of θ_c, $\rho_{\theta\tau}$, and $\rho_{b\beta}$. Each entry in Table 2 was calculated as a weighted average of the percentages of examinees who exceeded the time limit within each ability level, using weights proportional to the frequency of occurrence of each ability level under the assumption that $\theta \sim N(0, 1)$.

It is seen in Table 2 that across all conditions, the time-limited SPRT stopping rule was generally able to control the percentage of examinees who exceed the time limit to be below 5%, except under item selection method M1 when $\theta_c = 1$ and $\rho_{b\beta} = 0.65$. In addition, the percentage of examinees who exceeded the time limit was smaller under M2 than under M1 when both item selection methods were coupled with the time-limited SPRT stopping rule. This is consistent with the fact that tests administered under M2 required less time to be completed than those under M1. When the regular SPRT was used as a stopping rule in conjunction with item selection method M1, more than 40% of examinees took longer than 900 s to complete the tests in nearly all conditions studied. For this most commonly used combination of stopping rule and item selection method in CCT, the percentage of examinees whose testing time exceeded 900 s was lowest when $\theta_c = -1$ with $(\rho_{\theta\tau}, \rho_{b\beta}) = (0, 0.65)$ and highest when $\theta_c = 1$ with $(\rho_{\theta\tau}, \rho_{b\beta}) = (0.59, 0.65)$, but even the lowest percentage was higher than 18%.

Summary and Discussion

One purpose of this article is to use RT modeling to improve the maximum information item selection method commonly used in CCT. Instead of minimizing the number of administered items, improvement is sought by shortening the duration of the test without unduly compromising classification accuracy. Previous research (Fan et al., 2012) had demonstrated the success of the ''maximum information per time unit'' framework in reducing test duration in the context of CAT, where the goal is to estimate examinees' latent abilities. However, no previous research had applied the framework in the context of CCT, which is commonly used in many

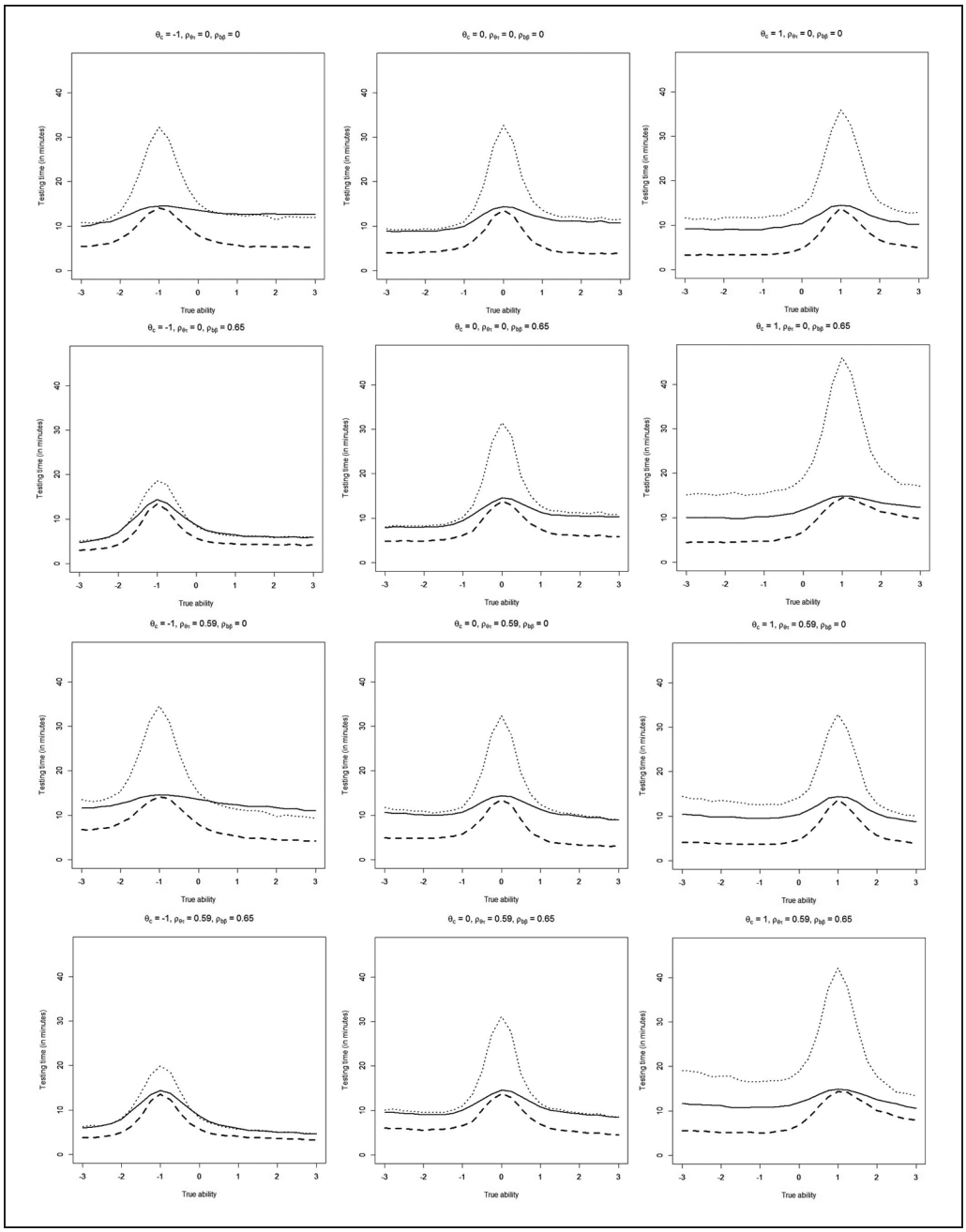

Figure 2. Time comparison between M1 under the regular SPRT (dotted line), M1 under the time-limited SPRT (solid line), and M2 under the time-limited SPRT (dashed line).

Note. Different columns represent different locations of the mastery threshold ($\theta_c = -1, 0, 1$ in the first, second, and third columns, respectively). Different rows represent different correlation structures (from the first to the fourth row: $\rho_{\theta\tau} = 0$ and $\rho_{b\beta} = 0$, $\rho_{\theta\tau} = 0$ and $\rho_{b\beta} = 0.65$, $\rho_{\theta\tau} = 0.59$ and $\rho_{b\beta} = 0$, $\rho_{\theta\tau} = 0.59$ and $\rho_{b\beta} = 0.65$). SPRT = Sequential Probability Ratio Test.

Table 2. Percentage of Examinees Who Exceeded a Time Limit of 900 s.

θ_c	$(\rho_{\theta\tau}, \rho_{b\beta})$	M1, regular SPRT	M1, time-limited SPRT	M2, time-limited SPRT
−1	(0, 0)	48.58	4.60	1.36
−1	(0, 0.65)	18.97	2.34	0.82
−1	(0.59, 0)	47.07	4.23	1.44
−1	(0.59, 0.65)	20.92	2.22	0.88
0	(0, 0)	48.62	4.90	1.75
0	(0, 0.65)	45.27	4.90	1.58
0	(0.59, 0)	47.79	4.79	1.67
0	(0.59, 0.65)	43.98	4.54	1.47
1	(0, 0)	45.41	3.94	1.17
1	(0, 0.65)	63.35	6.63	1.93
1	(0.59, 0)	45.65	3.28	0.88
1	(0.59, 0.65)	64.66	6.32	1.73

Note. SPRT = Sequential Probability Ratio Test.

certification or licensure tests as well as health and psychological assessments. Another purpose of this article is to use RT modeling to improve the SPRT stopping rule for CCT with a time constraint. The proposed stopping rule attempts to control the percentage of examinees who exceed the specified time limit for the test.

Results of the simulation study showed that taking into account the expected time spent by an examinee on an item substantially reduced testing time regardless of the location of the mastery threshold or the correlation structure between item parameters and between person parameters. By assessing the ATL, PCD, and average testing time, as well as the percentage of examinees who exceed the time limit, it was seen that the current standard in CCT that combines the item selection method M1 with the regular SPRT stopping rule is suboptimal from the perspective of testing time. An item selection method that considers testing time such as M2, coupled with a stopping rule that attempts to control the percentage of examinees who exceed a given time limit, may substantially reduce testing time without unduly sacrificing classification accuracy. In particular, the item selection method M2 not only reduced testing time but also generally had higher PCD than M1 when both were coupled with the time-limited SPRT stopping rule. Because the PCD of tests under item selection method M1 and the regular SPRT stopping rule was made as close as possible to those of tests where M1 was coupled with the time-limited SPRT stopping rule, the previous observation in turn implies that tests administered under the item selection method M2 coupled with the time-limited SPRT stopping rule had higher PCD than the standard practice in CCT.

Another simulation, the results of which are not presented herein, was performed to assess the performance of the method of Fan et al. (2012) in the current context of CCT. Noting that choosing test items that maximize information per time unit at an interim ability estimate combined with a fixed-information stopping rule is not the standard practice in CCT (nor is it designed for that testing situation), the results showed that the method resulted in tests with lower ATL as well as average testing time, and thus also a lower percentage of examinees who exceed the time limit.

As mentioned in the ‘‘Introduction’’ section, the definition of an efficient test is one that couples a high PCD with a short duration, not necessarily one with a small number of items. This makes M2 a suitable item selection method for CCT used in low-stakes educational tests or in health and psychological assessment, where there is little concern about item exposure. In these applications of CCT, it may be more important that the test be completed in a short period of

time, before the test-takers experience fatigue, become less focused or less motivated to complete the test. Moreover, in such contexts, it is in the test-takers' interest to find as much information as possible about themselves and thus they have little incentive to manipulate the system by simply waiting until the time limit is reached after correctly answering one item (after which their LR statistic is above the mastery threshold, when using $C = 1$).

In this article, the authors focused on evaluating the performance of the item selection methods as well as the proposed stopping rule in CCT with only one mastery threshold and two proficiency groups. With evidence suggesting the usefulness of choosing test items that maximize information per time unit to reduce testing time, as well as the usefulness of the stopping rule in controlling the percentage of examinees who exceed the given time limit, a possible direction for future research is to extend the methods to CCT with more than one mastery threshold. In addition, performance of the methods also needs to be further investigated under different structures of item pools, with different time limits, and under the presence of test constraints such as content balancing and item exposure control.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*, (pp. 397-472). Reading, MA: Addison-Wesley.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249-261.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*, 655-670.

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *33*, 442-463.

Lohr, K. N. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research*, *11*, 193-205.

Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computerized adaptive testing* (LSAC Computerized Testing Report No. 03–03). Newtown, PA: Law School Admission Council.

Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (ACT Research Report Series No. 93–7). Iowa City, IA: American College Testing.

Spray, J. A., & Reckase, M. D. (1994, April). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedure for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21*, 405-414.

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation, 16*(4). Retrieved from http://pareonline.net/getvn.asp?v=16&n=4

van der Linden, W. J. (1999). Empirical initialization of the ability estimator in adaptive testing. *Applied Psychological Measurement*, *23*, 21-29. [Erratum, 23, 248]

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5-20.

van der Linden, W. J. (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*, 25-41.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195-210.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika*, *68*, 251-265.

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*, 418-438.

Wald, A. (1947). *Sequential analysis*. New York, NY: John Wiley.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*, 41-58.